GENOMICS

Division of Microbiology and Infectious Diseases

The Division of Microbiology and Infectious Diseases supports a significant program in genomics research, including sequencing of human pathogens and invertebrate vectors of diseases, applying genomic and proteomic technologies to the study of microorganisms and infectious diseases, supporting genomic databases, and providing access to high-quality genomic resources and technologies to the scientific community.

Genome Sequencing

Advances in molecular biology have led to remarkably fast and accurate methods for sequencing the genomes of disease-causing microorganisms and invertebrate vectors of infectious diseases. Genome sequencing reveals the lineup of paired chemical bases that make up the pathogen's DNA, the language of life. When scientists identify genes that are unique to a particular microbe, drugs can be targeted to these genes, the products of these genes can be incorporated into experimental vaccines, and more specific and sensitive diagnostic tests can be developed. Furthermore, strategies can be devised to counteract genetic mutations that cause a microbe to become drug resistant. Once virulence genes are found, researchers can attempt to disable them. Moreover, genetic variations detected in different strains of the same pathogen can be used to study the population dynamics of these strains such as the spread of a virulent form of a pathogen in a susceptible population. Finally, understanding the genetic basis for both virulence and drug resistance also may help predict disease prognosis and influence the type and extent of patient care and treatment.

The potential payoffs of sequencing pathogens are enormous. NIAID scientists are exploiting sequence information in the following ways:

- To enhance the understanding of the disease-causing microorganisms and invertebrate vectors of diseases and their ability to cause disease;
- To facilitate the identification of novel and specific therapeutic targets for the development of new and improved drugs and vaccines, sequence-based detection technologies, and medical diagnostics for preventing, treating, and detecting infection;
- To compare the genomes of different species, strains, and clinical isolates to identify genetic variations and polymorphisms that may correlate with phenotypes such as drug resistance, infectivity, morbidity, and virulence; and to identify targets for improving both forensic strain identification and molecular genotyping of microorganisms; and
- To trace microbial evolution.

Recognizing the tremendous benefits of genome sequencing, NIAID has funded a significant number of projects to sequence the full genomes of medically important microbes, including the bacteria that cause anthrax, plague, tuberculosis, gonorrhea, chlamydia, cholera, strep throat, scarlet fever, and food-borne diseases. In total, NIAID-supported investigators have completed genome-sequencing projects for more than 37 bacteria, 5 parasitic protozoas, and 1 invertebrate vector of disease. NIAID's data release policies have ensured that the genome-sequencing data and their annotation are

available on publicly accessible Web sites, before publication, for immediate access and use by the scientific community.

NIAID collaborates with other funding agencies to sequence the larger genomes of protozoan pathogens such as the organism causing malaria and invertebrate vectors of infectious diseases. The complete genome sequence of *Plasmodium falciparum,* the parasite that causes malaria, was published in 2002, and is based on the work of the International Malaria Genome Sequencing Consortium, which NIAID supports. In addition, NIAID supported the rapid sequencing of the genome of the malaria mosquito, *Anopheles gambiae,* which transmits the malaria parasite to humans. The genome-sequencing project was completed and published in 2002. The three genome sequences—the malaria mosquito, malaria parasite, and the human—will provide scientists with a unique opportunity to study the natural history of malaria. For the first time, researchers will have the complete genetic information on an infectious organism, its natural host, and the insect that transmits the disease. NIAID-supported scientists already have taken advantage of this valuable genomic information and provided insights into the biology of the mosquito and parasite, molecular mechanisms involved in insecticide resistance, and gene and gene products that are potential targets for candidate drugs.

NIAID has made a significant investment in the DNA sequencing of the genomes of microorganisms considered agents of bioterrorism. NIAID currently is supporting sequencing efforts of several Category A, B, and C potential agents of bioterrorism. For example, with Institute support, the complete genome sequencing of *Bacillus anthracis* (Ames strain) has been completed in a collaborative effort with the Office of Naval Research (ONR) and the Department of Energy (DOE). Other organisms that have been sequenced include *Brucella suis, Burkholderia mallei, Clostridium perfringens, Coxiella burnetii,* and *Rickettsia typhi* with funds from the Defense Advanced Research Projects Agency (DARPA); *B. anthracis* (Kruger B1 and Western North America strains); *Staphylococcus aureus; Yersinia pestis; Mycobacterium tuberculosis (M.tb);* food-borne bacterial pathogens including diarrheagenic *Escherichia coli, Vibrio cholerae; Shigella, Salmonella;* and parasitic protozoa including *Cryptosporidium parvum, Giardia lamblia, Entamoeba histolytica,* and *Toxoplasma gondii.* In addition, NIAID has expanded its sequencing efforts of *B. anthracis* and has developed a comprehensive genomic analysis that includes sequencing of at least 14 additional strains, clinical isolates, near neighbors, and related species. Under this expansion, sequencing projects were completed for *B. anthracis* strains, Kruger B and Western North America strains, and a related species, *Bacillus cereus.* This important research will facilitate forensic strain identification; microbial pathogenesis; and the discovery of new targets for drugs, vaccines, and molecular signatures and biomarkers for diagnostics to combat bioterrorism.

Genomic Research

NIAID-supported investigators are applying genomic and emerging technologies to study microorganisms and infectious diseases. This activity includes both basic research such as studying the biology of the microorganism and host response to infection, and applied research such as development of medical

diagnostics, drugs, and vaccines. Genomic technologies are providing a new platform for scientists to study infectious agents at the whole genome or proteome level, providing clues to infectivity, pathogenesis, and virulence as well as the host response to infection, vaccines, and drugs.

- Whole genome and proteome expression studies are being used to identify pathogen-specific genes and proteins involved in virulence, pathogenesis, and disease transmission.
- Proteomic technologies are being applied to pathogen and/or host proteome characterization, and the identification of protein targets as potential candidates for the next generation of vaccines, therapeutics, and diagnostics.
- Genomic technologies are providing platforms for examining genetic variation in related species, strains, and clinical isolates and for studying host response to susceptibility to infection and effectiveness of vaccines and drugs.

A list of NIAID-supported large-scale pathogen genome-sequencing projects is provided on page 98.

Genomic Resources, Reagents, and Technologies

NIAID is committed to facilitating the access and distribution of genomic resources and technologies to the research community for functional genomic analysis of microbial pathogens and to supporting the development of bioinformatic and computational tools to allow investigators to store and manipulate genomic and postgenomic data.

NIAID supports a Pathogen Functional Genomics Resource Center (PFGRC) *(www.niaid.nih.gov/dmid/genomes/pfgrc/default.htm)* at the Institute of Genomic Research. PFGRC was established in 2001 to provide and distribute to the broader research community a wide range of genomic and related resources and technologies for the functional analysis of microbial pathogens and invertebrate vectors of infectious diseases. In fiscal year (FY) 2002, considerable progress was made toward this goal, including the selection and distribution to the research community of three organisms for generating microarrays, *Salmonella typhimurium*, *Staphylococcus aureus*, and *Streptococcus pneumoniae*, and the creation of an informative Internet-based Web site and Web-based application process. In addition, PFGRC was expanded to provide the research community with the needed resources and reagents to conduct both basic and applied research on microorganisms responsible for emerging and re-emerging infectious diseases and those considered agents of bioterrorism. Under this expansion in FY 2002, sequencing projects were completed for *B. anthracis* strains (Kruger B and Western North America strain) and a related species, *B. cereus.* In addition, Affymetrix-based microarray technology was added to PFGRC, and new genomic software tools have been developed for comparative genomics. In FY 2003, additional organism-specific microarrays were generated and distributed to the scientific community, including *M.tb, Neisseria gonorrhoeae*, and *P. falciparum.* PFGRC has provided severe acute respiratory syndrome (SARS) genomic resources to the broad scientific community to spur basic and applied research on SARS.

In FY 2003, NIAID awarded a contract to The Institute for Genome Research (TIGR) to support a Microbial Genome Sequencing

Focus on

NIAID Offers "SARS Chip" Free to Researchers

Severe acute respiratory syndrome (SARS) is a viral respiratory illness caused by the SARS-associated coronavirus (SARS-CoV). To spur research on SARS, NIAID initiated a program under which researchers can obtain at no cost a SARS-sequencing microarray. Also known as a "SARS chip," this device enables researchers to rapidly detect tiny genetic variations among different SARS virus strains. Access to this powerful tool will help scientists better understand the origin and spread of this newly recognized pathogen and will provide new leads in the search for effective SARS countermeasures. Distribution of the SARS arrays—GeneChip® brand made by Affymetrix, Inc.—will be coordinated by NIAID's Pathogen Functional Genomics Resource Center (PFGRC), which is operated under contract by The Institute for Genomic Research (TIGR). Any researcher can request the SARS array using a simple Web-based application process.

The new array was designed using data from research centers in the United States, Canada, and Asia that had sequenced the complete genome of the SARS coronavirus. The SARS array will help scientists achieve a number of objectives:

- By comparing the genomes of different SARS virus isolates, scientists can construct a "family tree" of the SARS coronaviruses;
- By comparing clinical outcomes among individuals infected with different SARS virus subtypes, scientists can determine which strains are most dangerous and gain key information for the development of targeted antiviral drugs;
- Over time, epidemiologists can trace how the virus evolves as it spreads to different populations in different geographic locales.

Details on the program are posted on the NIAID Web site at *www.niaid.nih.gov/dmid/genomes/pfgrc.* The SARS-CoV array program is part of an ongoing effort by NIAID to promote genomics as a vehicle to improve the understanding of disease-causing microbes and to develop new diagnostics, therapeutics, and vaccines to combat them.

Center to allow for rapid and cost-efficient production of high-quality, microbial genome sequences. Genomes to be sequenced include microorganisms considered agents of bioterrorism (NIAID Category A, B, and C), related organisms, clinical isolates, closely related species, and invertebrate vectors of infectious diseases and microorganisms responsible for emerging and re-emerging infectious diseases. NIAID's Microbial Genome Sequencing Center has the capacity to respond to national needs and Government priorities for genome sequencing, filling in sequence gaps and therefore providing genome sequencing data for multiple usages, including forensic strain identification and identifying targets for drugs, vaccines, and diagnostics.

In addition, the Malaria Research and Reference Reagent Resource (MR4) Center *(www.malaria.mr4.org)* continued to provide expanded access to quality-controlled reagents for the international malaria research community in 2002.

NIAID participates in a trans-NIH initiative, the Mammalian Full Length cDNA Project, known as the Mammalian Gene Collection (MGC) *(http://mgc.nci.nih.gov).* The goal of MGC is to generate a full set of human and other mammalian full-length (open reading frames) cDNAs and to support the production of cDNA libraries, clones, databases, and repositories that are accessible to the biomedical research community. Currently, the collection includes 12,414 human and 7,409 mouse full-length clones. These high-quality cDNA clones are a useful resource to the NIAID-supported scientific community for functional genomics and proteomics analysis of the host response to infection.

Bioinformatics and Databases

NIAID continues to provide support for databases of genomic and postgenomic information and analysis tools on sexually transmitted pathogens *(www.stdgen.lanl.gov)* and poxviruses *(www.poxvirus.org).* DARPA transferred funds to assist NIAID in supporting this poxviruses bioinformatics resource center *(www.poxvirus.org),* which includes a poxvirus genomic database, software for data analysis, a literature resource, and a repository of poxvirus species and strains, available at the American Type Culture Collection. These databases are a valuable genomic resource, providing the scientific community with easy access to large amounts of genomic and related data and bioinformatics tools for data analysis.

Genomics and Proteomics

NIAID continues to collaborate with the National Institute of General Medical Sciences (NIGMS) on the NIGMS Protein Structure Initiative *(www.nigms.nih.gov/funding/psi.html),* which supports research centers for the development of high-throughput methods and structural determination of proteins. One project supports the determination and analysis of structures of more than 400 functionally relevant *M.tb* proteins, and another project focuses on determining the protein structures from pathogenic protozoa. Structural and functional information on many proteins is now available in Web-based databases for access by the scientific community.

Division of Allergy, Immunology, and Transplantation

The Division of Allergy, Immunology, and Transplantation (DAIT) also supports genomics research. The human immune system is composed of complex networks of interacting cells, each programmed by precisely scripted genes. Underlying each immune response to a disease is a multistepped pathway of interacting molecules influenced by an individual's unique genomic characteristics. The immune system plays a critical role in diseases such as rheumatoid arthritis, hay fever, contact dermatitis, insulin-dependent or type 1 diabetes, systemic lupus erythematosus (SLE), and graft rejection of transplanted solid organs, tissues, and cells. Each of these diseases has an underlying genetic component.

Genomic research supported by DAIT is yielding insights into the functional and structural dimensions of immune system regulation, hypersensitivity, and inflammation in diseases such as asthma, the dysregulation of immune responses that results in autoimmune disease, and basic mechanisms of immune tolerance and graft rejection. This research is important in the following areas:

- **Asthma and allergic diseases.** DAIT-supported research on the genetics of asthma, hypersensitivity, inflammation, and T cell mediation enables us to understand the mechanisms underlying these immune responses, resulting in improved diagnostic, prevention, and treatment strategies. Through genomic research, DAIT-supported investigators discovered that interleukin-4 (IL-4), a cytokine that is produced by helper T cells and mast cells, stimulates antibody production by B cells in a series of reactions involving several genes. Further studies on IL-4 may provide a marker for measuring asthma risk and severity.

- **Autoimmune diseases.** DAIT supports research on type 1 diabetes and other autoimmune diseases that involve more than a single gene. Recent developments in genomics such as high-resolution DNA analysis and bioinformatics tools are making it possible to understand the underlying genetic causes of these complex diseases. For example, one approach compares the genes of individuals who have an autoimmune disease with those of healthy individuals to identify genetic and genomic differences that may be the underlying cause of disease. Between 10 and 20 distinct loci on the human genome may be responsible for susceptibility to type 1 diabetes. This knowledge will increase our ability to predict, diagnose, and treat this disease.

- **Transplantation.** DAIT-supported research on the genetics of graft rejection and immune tolerance is breaking new ground in the transplantation of solid organs, tissues, and cells for the prevention and treatment of disease. Genomic research funded by DAIT has identified surrogate markers of graft rejection in kidney transplant recipients. This research holds promise for the development of a noninvasive predictor of graft rejection based on gene expression analysis in urinary cells of transplant recipients.

- **Basic immunology research.** Basic research in immunology furthers our understanding of the properties, interactions, and functions of the cells of the immune system and the genetic aspects of immune system regulation and provides information about essential structural immunobiology. Recent breakthroughs in the basic science of immunogenetics inform clinical immunology, which may lead to the development of new immune-based therapies. Examples of basic immunology research supported by DAIT include the following:

  - Use of large-scale gene- and protein-expression analysis tools to describe pathways of cellular activation;

  - Discovery of anti-inflammatory and immunosuppressive agents using DNA-based screening methods; and

  - Analysis of genomic databases of T cell receptors and immunoglobulin gene sequences to link structural, functional, and clinical information.

Multicenter Research Programs

DAIT supports several multicenter research programs that include significant genomic efforts aimed at understanding the underlying mechanisms of immune-mediated diseases.

Immune Tolerance Network (ITN). The potential impact of tolerance induction to improve human health is great, encompassing a broad range of immune-mediated disorders, including autoimmune diseases such as type 1 diabetes, rheumatoid arthritis, and multiple

sclerosis; asthma and allergic diseases; and graft rejection in solid organ, tissue, and cell transplantation. With co-sponsorship from the National Institute of Diabetes and Digestive and Kidney Diseases and the Juvenile Diabetes Research Foundation International, NIAID supports ITN. ITN is an international consortium of more than 80 investigators in the United States, Canada, and Europe dedicated to the clinical evaluation of novel, tolerance-inducing therapies in autoimmune diseases; asthma and allergic diseases; and to preventing rejection of transplanted organs, tissues, and cells. The goal of these therapies is to re-educate the immune system to eliminate injurious immune responses and graft rejection while preserving protective immunity against infectious agents. ITN also conducts integrated studies on the underlying mechanisms of these approaches and develops and evaluates markers and assays to measure the induction, maintenance, and loss of tolerance in humans. Since its inception, ITN has initiated more than 15 clinical protocols, a variety of state-of-the-art core facilities, and about 5 additional studies designed to explore immune mechanisms leading to the development, maintenance, or loss of clinical tolerance.

Genomic research now under way in ITN may offer new therapeutic strategies for tolerance induction. ITN is developing clinical trials of multiple tolerance-induction approaches for several autoimmune diseases, including multiple sclerosis and type 1 diabetes. ITN also is pursuing clinical trials of multiple tolerance-induction approaches for asthma and allergic diseases, including a trial of DNA-ragweed-allergen conjugates for the treatment of allergic rhinitis. The network includes core laboratories to develop diagnostic assays to measure the induction, maintenance, and loss of tolerance in humans. These core facilities will develop and perform microarray analyses of gene expression, quantitative assays of T cell reactivity, novel tissue morphology studies to analyze tissue changes due to disease progression and therapeutic efficacy, and bioinformatics approaches to analyze clinical and scientific data sets from the ITN-sponsored clinical trials. More information on ITN is available on its Web site at *www.immunetolerance.org.*

Autoimmunity Centers of Excellence (ACEs). ACEs support collaborative basic and clinical research on autoimmune diseases, including single-site and multi-site pilot clinical trials of promising immunomodulatory therapies. ACEs presently are enrolling participants in several clinical trials, including a trial of anti-CD20 in SLE and a trial of anti-C5 in lupus nephritis.

International Histocompatibility Working Group (IHWG). IHWG is a network of more than 200 laboratories in more than 70 countries that applies new molecular techniques to population-based studies of histocompatibility genes. Histocompatibility genes allow the immune system to respond to specific pathogens, but these genes also play a role in the unwanted immune responses that occur in graft rejection and autoimmune diseases. Recent advances in genomics will facilitate the work of the human leukocyte antigen class II genes and related polymorphisms and their role in immunity, disease susceptibility, and graft rejection. Genomic techniques developed by IHWG investigators and others have shown a greater diversity among histocompatibility genes than was previously detected by conventional serologic methods. This work will bridge the

gap between serologic and genomic definitions of these genes.

Multiple Autoimmune Disease Genetics Consortium (MADGC). MADGC is a repository of genetic and clinical data and specimens from families in which two or more individuals are affected by two or more distinct autoimmune diseases. This resource provides materials to promote research aimed at discovering the human immune response genes involved in autoimmunity. MADGC began enrolling families in May 2000; to date, 162 families have been enrolled and 125 families are in the process, working toward the goal of 400 families enrolled by 2004. More information can be found at *www.madgc.org.*

North American Rheumatoid Arthritis Consortium (NARAC). NARAC is a collaborative registry and repository of information on families with rheumatoid arthritis. The NARAC database contains information on 902 families, encompassing 1,522 patient visits. Of the 902 families, data for more than half have been validated, including 600 affected sibling pairs. The family registry and the repository samples should facilitate the characterization of the genes underlying susceptibility to rheumatoid arthritis and are available to all investigators. More information can be found at *http://narac.patternrx.com.* This registry is co-sponsored by the National Institute of Arthritis and Musculoskeletal and Skin Diseases and the Arthritis Foundation.

Primary Immunodeficiency Diseases Registry and Consortium. The registry was established by NIAID through a contract with the Immune Deficiency Foundation to maintain clinical information on patients in the United States affected by primary immunodeficiency diseases. For each disease, the registry collects information on the natural course of the disease, including early and late complications, effects of therapy, and causes of death. The diseases included in the registry are chronic granulomatous disease, hyper-IgM syndrome, severe combined immunodeficiency disease disease, X-linked agammaglobulinemia, Wiskott-Aldrich syndrome, common variable immunodeficiency disease, leukocyte-adhesion deficiency, and DiGeorge syndrome. In FY 2003, the Primary Immunodeficiency Diseases Consortium was established. The Consortium will (1) provide leadership and mentoring; facilitate collaborations; enhance coordination of research efforts; and solicit, review, recommend, and make awards for pilot or small research projects; (2) maintain and expand a primary immunodeficiency diseases registry, which will provide data to the research community about the clinical characteristics and prevalence of these diseases; and (3) develop a repository of specimens from subjects with primary immunodeficiency diseases.

The following is a list of NIAID-supported large-scale pathogen genome-sequencing projects active in fiscal year 2003:

Organism	Disease
Aedes aegypt	yellow fever
Anopheles gambiae	malaria
Aspergillus fumigatus	aspergillosis
Brugia malayi	elephantiasis
Clostridium perfringens	gas gangrene
Coccidioides immitis	respiratory infections; coccidioidomycosis
Cryptococcus neoformans	cryptococcosis
Cryptosporidium parvum	food-borne and water-borne diseases, gastritis
Ehrlichia spp.	ehrlichiosis
Entamoeba histolytica	dysentery
Giardia lamblia	giardiasis
Histoplasma capsulatum	histoplasmosis
Legionella pneumophila	Legionnaire's disease
Leishmania major	cutaneous leishmaniasis
Nematode species	helminthiases
Plasmodium vivax	malaria
Pneumocystis carinii	pneumonia, opportunistic disease
Rickettsia rickettsii	Rocky Mountain spotted fever
Rickettsia typhi	typhus
Salmonella typhi	typhoid fever
Schistosoma mansoni	dermatitis, Katayama fever, liver inflammation, fibrosis
Streptococcus agalactiae	Group B Streptococcus
Streptococcus pneumoniae	pneumonia, meningitis
Toxoplasma gondii	toxoplasmosis, congenital, and ocular infections, opportunistic disease
Trichomonas vaginalis	vaginitis
Trypanosoma brucei	trypanosomiasis
Trypanosoma cruzi	Chagas' disease
Wolbachia	endosymbiont of filarial nematodes and insect vectors